HAL
open science

The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines

Jon Barker, Shinji Watanabe, Emmanuel Vincent, Jan Trmal

To cite this version:

Jon Barker, Shinji Watanabe, Emmanuel Vincent, Jan Trmal. The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines. Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association, Sep 2018, Hyderabad, India. hal-01744021

HAL Id: hal-01744021

https://inria.hal.science/hal-01744021v1

Submitted on 27 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines

Jon Barker[1], *Shinji Watanabe*[2], *Emmanuel Vincent*[3], *and Jan Trmal*[2]

[1]University of Sheffield, UK
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA
[3]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

j.p.barker@sheffield.ac.uk, shinjiw@jhu.edu, emmanuel.vincent@inria.fr, jtrmal@gmail.com

Abstract

The CHiME challenge series aims to advance robust automatic speech recognition (ASR) technology by promoting research at the interface of speech and language processing, signal processing, and machine learning. This paper introduces the 5th CHiME Challenge, which considers the task of distant multi-microphone conversational ASR in real home environments. Speech material was elicited using a dinner party scenario with efforts taken to capture data that is representative of natural conversational speech and recorded by 6 Kinect microphone arrays and 4 binaural microphone pairs. The challenge features a single-array track and a multiple-array track and, for each track, distinct rankings will be produced for systems focusing on robustness with respect to distant-microphone capture vs. systems attempting to address all aspects of the task including conversational language modeling. We discuss the rationale for the challenge and provide a detailed description of the data collection procedure, the task, and the baseline systems for array synchronization, speech enhancement, and conventional and end-to-end ASR.

Index Terms: Robust ASR, noise, reverberation, conversational speech, microphone array, 'CHiME' challenge.

1. Introduction

Automatic speech recognition (ASR) performance in difficult reverberant and noisy conditions has improved tremendously in the last decade [1–5]. This can be attributed to advances in speech processing, audio enhancement, and machine learning, but also to the availability of real speech corpora recorded in cars [6, 7], quiet indoor environments [8, 9], noisy indoor and outdoor environments [10, 11], and challenging broadcast media [12, 13]. Among the applications of robust ASR, voice command in domestic environments has attracted much interest recently, due in particular to the release of Amazon Echo, Google Home and other devices targeting home automation and multimedia systems. The CHiME-1 [14] and CHiME-2 [15] challenges and corpora have contributed to popularizing research on this topic, together with the DICIT [16], Sweet-Home [17], and DIRHA [18] corpora. These corpora feature single-speaker reverberant and/or noisy speech recorded or simulated in a single home, which precludes the use of modern speech enhancement techniques based on machine learning. The recently released voiceHome corpus [19] addresses this issue, but the amount of data remains fairly small.

In parallel to research on acoustic robustness, research on conversational speech recognition has also made great progress, as illustrated by the recent announcements of super-human performance [20, 21] achieved on the Switchboard telephone conversation task [22] and by the ASpIRE challenge [23]. Distant-microphone recognition of noisy, overlapping, conversational speech is now widely believed to be the next frontier. Early attempts in this direction can be traced back to the ICSI [24], CHIL [25], and AMI [26] meeting corpora, the LLSEC [27] and COSINE [28] face-to-face interaction corpora, and the Sheffield Wargames corpus [29]. These corpora were recorded using advanced microphone array prototypes which are not commercially available, and as result could only be installed in a few laboratory rooms. The Santa Barbara Corpus of Spoken American English [30] stands out as the only large-scale corpus of naturally occurring spoken interactions between a wide variety of people recorded in real everyday situations including face-to-face or telephone conversations, card games, food preparation, on-the-job talk, story-telling, and more. Unfortunately, it was recorded via a single microphone.

The CHiME-5 Challenge aims to bridge the gap between these attempts by providing the first large-scale corpus of real multi-speaker conversational speech recorded via commercially available multi-microphone hardware in multiple homes. Speech material was elicited using a 4-people dinner party scenario and recorded by 6 distant Kinect microphone arrays and 4 binaural microphone pairs in 20 homes. The challenge features a single-array track and a multiple-array track. Distinct rankings will be produced for systems focusing on acoustic robustness vs. systems aiming to address all aspects of the task.

The paper is structured as follows. Sections 2 and 3 describe the data collection procedure and the task to be solved. Section 4 presents the baseline systems for array synchronization, speech enhancement, and ASR and the corresponding results. We conclude in Section 5.

2. Dataset

2.1. The scenario

The dataset is made up of the recording of twenty separate dinner parties taking place in real homes. Each dinner party has four participants - two acting as hosts and two as guests. The party members are all friends who know each other well and who are instructed to behave naturally.

Efforts have been taken to make the parties as natural as possible. The only constraints are that each party should last a minimum of 2 hours and should be composed of three phases, each corresponding to a different location: i) *kitchen* – preparing the meal in the kitchen area; ii) *dining* – eating the meal in the dining area; iii) *living* – a post-dinner period in a separate living room area.

Participants were allowed to move naturally from one loca-

tion to another but with the instruction that each phase should last at least 30 minutes. Participants were left free to converse on any topics of their choosing. Some personally identifying material was redacted post-recording as part of the consent process. Background television and commercial music were disallowed in order to avoid capturing copyrighted content.

2.2. Audio

Each party has been recorded with a set of six Microsoft Kinect devices. The devices have been strategically placed such that there are always at least two capturing the activity in each location. Each Kinect device has a linear array of 4 sample-synchronised microphones and a camera. The raw microphone signals and video have been recorded. Each Kinect is recorded onto a separate laptop computer. Floor plans were drafted to record the layout of the living space and the approximate location and orientation of each Kinect device.

In addition to the Kinects, to facilitate transcription, each participant wore a set of Soundman OKM II Classic Studio binaural microphones. The audio from these was recorded via a Soundman A3 adapter onto Tascam DR-05 stereo recorders also being worn by the participants.

2.3. Transcriptions

The parties have been fully transcribed. For each speaker a reference transcription is constructed in which, for each utterance produced by that speaker, the start and end times and the word sequence are manually obtained by listening to the speaker's binaural recording (the reference signal). For each other recording device, the utterance's start and end time are produced by shifting the reference timings by an amount that compensates for the a synchrony between devices (see Section 4.1).

The transcriptions can also contain the followings tags: $[noise]$ denoting any non-language noise made by the speaker (e.g., coughing, loud chewing, etc.); $[inaudible]$ denoting speech that is not clear enough to be transcribed; $[laughs]$ denoting instances of laughter; $[redacted]$ are parts of the signals that have been zeroed out for privacy reasons.

3. Task

3.1. Training, development, and evaluation sets

The 20 parties have been divided into disjoint training, development and evaluation sets as summarised in Table 1. There is no overlap between the speakers in each set.

Table 1: *Overview of CHiME-5 datasets*

Dataset	Parties	Speakers	Hours	Utterances
Train	16	32	40:33	79,980
Dev	2	8	4:27	7,440
Eval	2	8	5:12	11,028

For the development and evaluation data, the transcription file also contains a speaker location and 'reference' array for each utterance. The location can be either 'kitchen','dining room', or 'living room' and the reference array (the target for speech recognition) is chosen to be one that is situated in the same area.

3.2. Tracks and ranking

The challenge features two tracks:

- *single-array*: only the reference array can be used to recognise a given evaluation utterance,
- *multiple-array*: all arrays can be used.

For each track, two separate rankings will be produced:

- *Ranking A* – systems based on conventional acoustic modeling and using the supplied official language model: the outputs of the acoustic model must remain frame-level tied phonetic (senone) targets and the lexicon and language model must not be modified,
- *Ranking B* – all other systems, e.g., including systems based on end-to-end processing or systems whose lexicon and/or language model have been modified.

In other words, ranking A focuses on acoustic robustness only, while ranking B addresses all aspects of the task.

3.3. Instructions

A set of instructions has been provided that ensure that systems are broadly comparable and that participants respect the application scenario. In particular, systems are allowed to exploit knowledge of the utterance start and end time, the utterance speaker label and the speaker location label. During evaluation participants can use the entire session recording from the reference array (for the single-array track) or from all arrays (for multiple-array track), i.e., one can use the past and future acoustic context surrounding the utterance to be recognised. For training and development, participants are also provided with the binaural microphone signals and the floor plans. Participants are forbidden from manual modification of the data or the annotations (e.g., manual refinement of the utterance timings or transcriptions).

It is required that all parameters are tuned on the training set or development set. Participants can evaluate different versions of their system but the final submission will be the one that performs best on the development data, and this will be ranked according to its performance on the evaluation data. While, some modifications of the development set are necessarily allowed (e.g. automatic signal enhancement, or refinement of utterance timings), participants have been cautioned against techniques designed to fit the development data to the evaluation data (e.g. by selecting subsets, or systematically varying its pitch or level). These "biased" transformations are forbidden.

The challenge has been designed to promote research covering all stages in the recognition pipeline. Hence, participants are free to replace or improve any component of the baseline system, or even to replace the entire baseline with their own systems. However, the architecture of the system will determine whether a participant's result is ranked in category A or category B (see Section 3.2).

Participants will evaluate their own systems and will be asked to return overall WERs for the development and evaluation data, plus WERs broken down by session and location. They will also be asked to submit the corresponding lattices in Kaldi format to allow their scores to be validated, plus a technical description of their system.

4. Baselines

4.1. Array synchronization

While signals recorded by the same device are sample-synchronous, there is no precise synchronisation between devices. Across devices synchronisation cannot be guaranteed. The signal start times are approximately synchronised post-recording using a synchronisation tone that was played at the beginning of each recording session. However, devices can drift out of synchrony due to small variations in clock speed (clock drift) and due to frame dropping. To correct for this a cross-correlation approach is used to estimate the delay between one of the binaural recorders chosen as the reference and all other devices [31]. These delays are estimated at regular 10 second intervals throughout the recording. Using the delay estimates, separate utterance start and end times have been computed for each device and are recorded in the JSON transcription files.

4.2. Speech enhancement

CHiME-5 uses a weighted delay-and-sum beamformer (BeamformIt [32]) as a default multichannel speech enhancement approach, similar to the CHiME-4 recipe [11]. Beamforming is performed by using four microphone signals attached to the reference array. The reference array information is provided by the organizers through the JSON transcription file.

4.3. Conventional ASR

The conventional ASR baseline is distributed through the Kaldi github repository [33][1] and is described in brief below.

4.3.1. Data preparation (stage 0 and 1)

These stages provide Kaldi-format data directories, lexicons, and language models. We use a CMU dictionary[2] as a basic pronunciation dictionary. However, since the CHiME-5 conversations are spontaneous speech and a number of words are not present in the CMU dictionary, we use grapheme to phoneme conversion based on Phonetisaurus G2P [34][3] to provide the pronunciations of these OOV (out-of-vocabulary) words. The language model is selected automatically, based on perplexity on training data, but at the time of the writing, the selected LM is 3-gram trained by the MaxEnt modeling method as implemented in the SRILM toolkit [35–37]. The total vocabulary size is 128K augmented by the G2P process mentioned above.

4.3.2. Enhancement (stage 2)

This stage calls BeamformIt based speech enhancement, as introduced in Section 4.2.

4.3.3. Feature extraction and data arrangement (stage 3-6)

These stages include MFCC-based feature extraction for GMM training, and training data preparation (250k utterances, in `data/train_worn_u100k`) The training data combines both left and right channels (150k utterances) of the binaural microphone data (`data/train_worn`) and a subset (100k utterances) of all Kinect microphone data (`data/train_u100k`). Note that we observed some performance improvements when we use larger amounts of training data instead of the above subset. However, we have limited the size of the data in the baseline so that experiments can be run without requiring unreasonable computational resources.

4.3.4. HMM/GMM (stage 7-16)

Training and recognition are performed with a hidden Markov model (HMM) / Gaussian mixture model (GMM) system. The GMM stages include standard triphone-based acoustic model building with various feature transformations including linear discriminant analysis (LDA), maximum likelihood linear transformation (MLLT), and feature space maximum likelihood linear regression (fMLLR) with speaker adaptive training (SAT).

4.3.5. Data cleanup (stage 17)

This stage removes several irregular utterances, which improves the final performance of the system [38]. Totally 15% of utterances in the training data are excluded due to this cleaning process, which yields consistent improvement in the following LF-MMI TDNN training.

4.3.6. LF-MMI TDNN (stage 18)

This is an advanced time-delayed neural network (TDNN) baseline using lattice-free maximum mutual information (LF-MMI) training [39]. This baseline requires much larger computational resources: multiple GPUs for TDNN training (18 hours with 2-4 GPUs), many CPUs for i-vector and lattice generation, and large storage space for data augmentation (speed perturbation).

As a summary, compared with the previous CHiME-4 baseline [11], the CHiME-5 baseline introduces: 1) grapheme to phoneme conversion; 2) Data cleaning up; 3) Lattice free MMI training. With these techniques, we can provide a reasonable ASR baseline for this challenging task.

4.4. End-to-end ASR

CHiME-5 also provides an end-to-end ASR baseline based on ESPnet[4], which uses Chainer [40] and PyTorch [41], as its underlying deep learning engine.

4.4.1. Data preparation (stage 0)

This is the same as the Kaldi data directory preparation, as discussed in Section 4.3.1. However, the end-to-end ASR baseline does not require lexicon generation and FST preparation. This stage also includes beamforming based on the BeamformIt toolkit, as introduced in Section 4.2.

4.4.2. Feature extraction (stage 1)

This stage use the Kaldi feature extraction to generate Log-Mel-filterbank and pitch features (totally 83 dimensions). It also provides training data preparation (350k utterances, in `data/train_worn_u200k`), which combines both left and right channels (150k utterances) of the binaural microphone data (`data/train_worn`) and a subset (200k utterances) of all Kinect microphone data (`data/train_u200k`).

4.4.3. Data conversion for ESPnet (stage 2)

This stage converts all the information included in the Kaldi data directory (transcriptions, speaker IDs, and input and out-

[1] `https://github.com/kaldi-asr/kaldi/tree/master/egs/chime5/s5`

[2] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

[3] `https://github.com/AdolfVonKleist/Phonetisaurus`

[4] `https://github.com/espnet/espnet`

put lengths) to one JSON file (`data.json`) except for input features. This stage also creates a character table (45 characters appeared in the transcriptions).

4.4.4. Language model training (stage 3)

Character-based LSTM language model is trained by using either a Chainer or PyTorch backend, which is integrated with a decoder network in the following recognition stage.

4.4.5. End-to-end model training (stage 4)

A hybrid CTC/attention-based encoder-decoder network [42] is trained by using either the Chainer or PyTorch backend. The total training time is 12 hours with a single GPU (TitanX) when we use the PyTorch backend, which is less than the computational resources required for the Kaldi LF-MMI TDNN training (18 hours with 2-4 GPUs).

4.4.6. Recognition (stage 5)

Speech recognition is performed by combining the LSTM language model and end-to-end ASR model trained by previous stages with multiple CPUs.

4.5. Baseline results

Tables 2 and 3 provide the word error rates (WERs) of the binaural (oracle) and reference Kinect array (challenge baseline) microphones. The WERs of the challenge baseline are quite high due to very challenging environments of CHiME-5 for all of methods[5]. Comparing these tables, there is a significant performance difference between the array and binaural microphone results (e.g., 33.4% absolutely in LF-MMI TDNN), which indicates that the main difficulty of this challenge comes from the source and microphone distance in addition to the spontaneous and overlapped nature of the speech, which exist in both array and binaural microphone conditions. So, a major part of the challenge lies in developing speech enhancement techniques that can improve the challenge baseline to the level of the binaural microphone performance.

Table 2: *WERs for the development set using the binaural microphones (oracle).*

	Development set
conventional (GMM)	72.8
conventional (LF-MMI TDNN)	47.9
end-to-end	67.2

Table 3: *WERs for the development set using the reference Kinect array with beamforming (challenge baseline).*

	Development set
conventional (GMM)	91.7
conventional (LF-MMI TDNN)	81.3
end-to-end	94.7

[5]Note, the current end-to-end ASR baseline performs poorly due to an insufficient amount of training data. However, the result of end-to-end ASR was better than that of the Kaldi GMM system when we used the binaural microphones for testing, which shows end-to-end ASR to be a promising direction for this challenging environment.

Table 4 shows the WER of the LF-MMI TDNN system with the development set for each session and room. The challenge participants have to submit this form with the evaluation set. We observe that performance is poorest in the kitchen condition, probably due to the kitchen background noises and greater degree of speaker movement that occurs in this location.

Table 4: *WERs of the LF-MMI TDNN system for each session and room conditions. The challenge participants have to submit this form scored with the evaluation set.*

	Development set	
	S02	S09
KITCHEN	87.3	81.6
DINING	79.5	80.6
LIVING	79.0	77.6

5. Conclusion

The 'CHiME' challenge series is aimed at evaluating ASR in real-world conditions. This paper has presented the 5th edition which targets conversational speech in an informal dinner party scenario recorded with multiple microphone arrays. The full dataset and state-of-the-art software baselines have been made publicly available. A set of challenge instructions has been carefully designed to allow meaningful comparison between systems and maximise scientific outcomes. The submitted systems and the results will be announced at the 5th 'CHiME' ISCA Workshop.

6. Acknowledgements

We would like to thank Google for funding the full data collection and annotation, Microsoft Research for providing Kinects, and Microsoft India for sponsoring the 5th 'CHiME' Workshop. E. Vincent acknowledges support from the French National Research Agency in the framework of the project VOCADOM "Robust voice command adapted to the user and to the context for AAL" (ANR-16-CE33-0006).

7. References

[1] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[2] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition — A Bridge to Practical Applications*. Elsevier, 2015.

[3] S. Watanabe, M. Delcroix, F. Metze, and J. R. Hershey, Eds., *New Era for Robust Speech Recognition — Exploiting Deep Learning*. Springer, 2017.

[4] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[5] S. Makino, Ed., *Audio Source Separation*. Springer, 2018.

[6] http://aurora.hsnr.de/aurora-3/reports.html.

[7] J. H. L. Hansen, P. Angkititrakul, J. Plucienkowski, S. Gallant, U. Yapanel, B. Pellom, W. Ward, and R. Cole, ""CU-Move": Analysis & corpus development for interactive in-vehicle speech systems," in *Proc. Eurospeech*, 2001, pp. 2023–2026.

[8] L. Lamel, F. Schiel, A. Fourcin, J. Mariani, and H. Tillman, "The translingual English database (TED)," in *Proc. 3rd Int. Conf. on Spoken Language Processing (ICSLP)*, 1994.

[9] E. Zwyssig, F. Faubel, S. Renals, and M. Lincoln, "Recognition of overlapping speech using digital MEMS microphone arrays," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7068–7072.

[10] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHIME' speech separation and recognition challenge: Analysis and outcomes," *Computer Speech and Language*, vol. 46, pp. 605–626, 2017.

[11] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, vol. 46, pp. 535–557, 2017.

[12] G. Gravier, G. Adda, N. Paulsson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language," in *Proc. 8th Int. Conf. on Language Resources and Evaluation (LREC)*, 2012, pp. 114–118.

[13] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. C. Woodland, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 687–693.

[14] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Comp. Speech and Lang.*, vol. 27, no. 3, pp. 621–633, May 2013.

[15] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: An overview of challenge systems and outcomes," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013, pp. 162–167.

[16] A. Brutti, L. Cristoforetti, W. Kellermann, L. Marquardt, and M. Omologo, "WOZ acoustic data collection for interactive TV," in *Proc. 6th Int. Conf. on Language Resources and Evaluation (LREC)*, 2008, pp. 2330–2334.

[17] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, "The Sweet-Home speech and multimodal corpus for home automation interaction," in *Proc. 9th Int. Conf. on Language Resources and Evaluation (LREC)*, 2014, pp. 4499–4509.

[18] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 275–282.

[19] N. Bertin, E. Camberlein, E. Vincent, R. Lebarbenchon, S. Peillon, É. Lamandé, S. Sivasankaran, F. Bimbot, I. Illina, A. Tom, S. Fleury, and E. Jamet, "A French corpus for distant-microphone speech processing in real homes," in *Proc. Interspeech*, 2016, pp. 2781–2785.

[20] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," arXiv:1610.05256, 2017.

[21] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English conversational telephone speech recognition by humans and machines," arXiv:1703.02136, 2017.

[22] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. IEEE International Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, vol. 1, 1992, pp. 517–520.

[23] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 547–554.

[24] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003, pp. 364–367.

[25] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. Chu, A. Tyagi, J. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Language Resources and Evaluation*, vol. 41, no. 3–4, pp. 389–407, 2007.

[26] S. Renals, T. Hain, and H. Bourlard, "Interpretation of multiparty meetings: The AMI and AMIDA projects," in *Proc. 2nd Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2008, pp. 115–118.

[27] https://www.ll.mit.edu/mission/cybersec/HLT/corpora/SpeechCorpora.html.

[28] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, and J. Bilmes, "The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments," *Computer Speech and Language*, vol. 26, no. 1, pp. 52–66, 2011.

[29] C. Fox, Y. Liu, E. Zwyssig, and T. Hain, "The Sheffield wargames corpus," in *Proc. Interspeech*, 2013, pp. 1116–1120.

[30] J. W. Du Bois, W. L. Chafe, C. Meyer, S. A. Thompson, R. Englebretson, and N. Martey, "Santa Barbara corpus of spoken American English, parts 1–4," Linguistic Data Consortium.

[31] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[32] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 15, no. 7, pp. 2011–2023, 2007.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[34] J. R. Novak, N. Minematsu, and K. Hirose, "WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding," in *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, 2012, pp. 45–49.

[35] J. Wu and S. Khudanpur, "Building a topic-dependent maximum entropy model for very large corpora," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2002, pp. I–777–I–780.

[36] T. Alumäe and M. Kurimo, "Efficient estimation of maximum entropy language models with N-gram features: an SRILM extension," in *Proc. Interspeech*, Chiba, Japan, September 2010.

[37] A. Stolcke *et al.*, "SRILM-an extensible language modeling toolkit." in *Interspeech*, vol. 2002, 2002, pp. 901–904. [Online]. Available: http://www.speech.sri.com/projects/srilm/

[38] V. Peddinti, V. Manohar, Y. Wang, D. Povey, and S. Khudanpur, "Far-field ASR without parallel data." in *INTERSPEECH*, 2016, pp. 1996–2000.

[39] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, 2016, pp. 2751–2755.

[40] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a next-generation open source framework for deep learning," in *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, vol. 5, 2015.

[41] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proceedings of The future of gradient-based machine learning software and techniques (Autodiff) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, 2017.

[42] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.